The role of large pre-trained models in ecology and biodiversity conservation: Opportunities and Challenges

Hideyuki Doi[1], Takeshi Osawa[2], and Narumasa Tsutsumida[3]

[1]Kyoto University
[2]National Institute for Agro-environmental Science
[3]Saitama University

March 25, 2024

Abstract

Large pre-trained models (LPMs) have the capabilities to understand natural language, code, and diverse data including images; e.g., large language models (LLMs), code-generative models, and large vision models (LVMs) as well as combined as multi-modal models. We outlines the main applications of LPMs and multi-modal LPMs for ecology and biodiversity conservation. These applications include generating ecological data, generating code, providing insights into public opinion and sentiment. We highlighted the significant potential of LPMs and the potential use of Ecology-specialized LPMs for ecology and biodiversity conservation. They offer unprecedented opportunities for analyzing diverse data, extracting meaningful insights, and informing conservation decisions.

Viewpoint manuscript for Ecology Letters

The role of large pre-trained models in ecology and biodiversity conservation: Opportunities and Challenges

Hideyuki Doi[1,*,+], Takeshi Osawa[2,*,+], and Narumasa Tsutsumida[3,*,+]

[1] Graduate School of Informatics, Kyoto University, Yoshida-honmachi, Sakyo-ku, Kyoto, 606-8501, Japan

[2] Graduate School of Urban Environmental Sciences, Tokyo Metropolitan University, Minami-Osawa 1-1, Hachiouji, Tokyo, 192-0397, Japan

[3] Graduate School of Science & Engineering, Saitama University, 255 Shimo-Okubo, Sakura ward, Saitama City, Saitama, 338-8570, Japan

+ These authors equally contributed to this study and are ordered alphabetically.

*Corresponding authors:

Hideyuki Doi, hideyuki.doi@icloud.com

Takeshi Osawa, arosawa@gmail.com

Narumasa Tsutsumida, rsnaru.jp@gmail.com

Authorship statement

All authors contributed to the conceptualization of this viewpoint. HD wrote the first draft. all authors contributed to writing, review, and editing.

A data statement

We did not use any data in this Viewpoint paper.

Short running title

LPMs in ecology and conservation

Keywords

artificial intelligence, modeling, digital data, public opinion and sentiment, remote sensing, biodiversity

Type of article: Viewpoint

Number of words in abstract: 100

Number of words in main text: 2041

Number of cited references: 25

Number of figures: 1

Number of tables: 0

Abstract

Large pre-trained models (LPMs) have the capabilities to understand natural language, code, and diverse data including images; e.g., large language models (LLMs), code-generative models, and large vision models (LVMs) as well as combined as multi-modal models. We outlines the main applications of LPMs and multi-modal LPMs for ecology and biodiversity conservation. These applications include generating ecological data, generating code, providing insights into public opinion and sentiment. We highlighted the significant potential of LPMs and the potential use of Ecology-specialized LPMs for ecology and biodiversity conservation. They offer unprecedented opportunities for analyzing diverse data, extracting meaningful insights, and informing conservation decisions.

Recent advancements in artificial intelligence (AI) have led to the emergence of sophisticated large pre-trained models (LPMs). These LPMs represent a significant leap forward in AI technology, offering unparalleled capabilities in understanding natural language, generating human-like text, and performing complex tasks across various domains. LPMs have shown impressive proficiency in various applications, including natural language processing, code generation, and data/vision analysis.

For natural language processing, large language models (LLMs) like Generative Pre-trained Transformer (GPT) (Brown *et al.* 2020), such as ChatGPT (GPT-3 and 4), Llama, and Bard, leading the revolution in natural language processing. These models possess the ability to generate human-like text, decipher complex language structures, and even translate languages with remarkable accuracy. LLMs prove invaluable for sentiment analysis, text summarization, and cross-language communication (e.g., Rane *et al.* 2023; Patil *et al.*2024a). Also, code-generative LLM, like GitHub Copilot (https://github.com/features/copilot) and CodeT5 (https://github.com/salesforce/CodeT5), contribute to the programming of the codes for analysis and simulation by understanding complex code and aiding developers in debugging. Large vision models (LVMs) have the ability to process and interpret vast amounts of visual data with exceptional accuracy and efficiency. LVMs utilize deep learning techniques and extensive pre-training on massive image datasets.

Multi-modal LPMs is a combined model that integrates diverse data types, such as text, images, and audio, broadening their scope. Using LLM+LVM, the most developed multi-modal LPM, text-to-image (T2I) synthesis has undergone significant advancements, particularly with the emergence of LLM and their enhancement in LVM (Cheng *et al.* 2024). These models excel in tasks like image captioning, visual question answering, and audio transcription, enabling more context-rich AI systems. The potential applications of

LPMs span from chatbots and virtual assistants to content creation and language learning. A significant advantage lies in their capacity to produce text indistinguishable from human writing, paving the way for more natural interactions with chatbots and virtual assistants (e.g., Rane *et al.* 2023 in medical). Furthermore, their language translation prowess holds the potential to reshape global communication (Patil*et al.* 2024a). Therefore, LPMs are now emerging as promising tools to tackle ecological challenges by analyzing vast data extracting meaningful insights, and forming a foundation for well-informed conservation decisions. The integration of LPMs in biodiversity conservation holds immense potential to revolutionize research practices and bolster conservation efforts. The subsequent sections delineate the applications of LPMs in ecology and biodiversity conservation.

Ecological data generation by LPMs

LPMs play a pivotal role in processing extensive documents on ecology, including scientific papers, reports, and online news articles, extracting pertinent information. For instance, researchers leverage LLMs to extract the description to identify novel or endangered species, monitor changes in population size, and discern emerging threats to biodiversity (e.g., Fabian et al. 2024). Many LLMs exhibit the ability to translate across languages, enabling researchers to access crucial information from non-English resources. Some studies suggested the potential of non-English resources in filling conservation and ecological knowledge gaps (e.g., Amano *et al.* 2021; Hannah*et al.* 2024). Despite biodiversity information being foundational to ecology, evolution, and conservation, Amano (*et al.* 2023) reported that non-English-language literature constituted 65% of the references cited as biodiversity information. Therefore, numerous natural history resources in non-English languages remain unshared due to the language barriers. LLMs facilitate the integration of natural history knowledge across languages, mitigating obstacles for the researchers in selecting the language for publication.

LPMs excel in analyzing the data, such as satellite information, to formulate statistical models that elucidate intricate relationships between species and their surroundings by code generation. A multi-modal LPM, LLM-LVM, can generate text from image analysis by automatically recognizing and classifying species names based on visual data (e.g., Jain *et al.* 2023; Parashar*et al.* 2023). This feature can enhance biodiversity monitoring and species identification efforts, particularly in large-scale surveys (e.g., Maurer *et al.* 2020; Wang *et al.* 2020). Therefore, LLM-LVM can process satellite imagery and extract valuable environmental data, such as land cover types, vegetation indices, and habitat connectivity. The combined use of text mining, image analysis, and satellite data processing by LLM-LVM has great potential for generating ecological data, which can enable researchers and conservation practitioners to gain valuable insights for effective biodiversity conservation and management strategies.

Code-generative LLMs foe ecolgical modeling and simulation

Code-generative LLMs play a crucial role in simulation research by promoting reproducibility and interpretability, ensuring that research findings can be validated and comprehended by the broader scientific community. For example, PyRates (Gast *et al.* 2023), a code-generation tool for dynamical systems modeling applied to biological systems, can be used for the dynamical models on ecological insights and ecosystem management. Therefore, LLMs present a unique opportunity for the standardization and commoditization of various research aspects, streamlining processes from SQL and GBIF data to R or Python codes. Researchers, by harnessing LLMs, can redirect their focus toward leveraging the rich and diverse field data accessible through platforms like GBIF, fostering a more comprehensive understanding of biodiversity patterns and trends (Anderson et al. 2023). A research approach with code-generative LLMs establishes a robust foundation for integrating logical support and employing case studies to address intricate conservation challenges, providing a radical yet promising avenue to advance biodiversity understanding and guide effective conservation strategies.

However, the utilization of code-generative LLMs in ecological modeling poses potential limitations and challenges. A major concern revolves around the risk of bias in the generated code, potentially leading to inaccurate or misleading predictions (Shah *et al.* 2020; Weidinger*et al.* 2021; Albrecht *et al.* 2022). LLMs,

being trained on large datasets, may inherently contain biases and limitations that are not always apparent to users. Additionally, LLMs may generate code that proves challenging to understand or modify, restricting the flexibility and adaptability of ecological models. To surmount these challenges, it is imperative to rigorously evaluate the performance and accuracy of LLM-generated code in ecological modeling applications.

Providing insights into public opinion and sentiment

LPMs have the potential to provide valuable insights into public opinion and sentiment regarding biodiversity conservation. One application involves using AI to generate summaries of complex environmental reports, making them easier for the general public to understand. This approach could enhance communication and promote wider engagement with ecological and conservation issues.

LLMs can analyze a wide range of sources, including social media data, which has increasingly become a significant platform for public discourse. For instance, Lee *et al.* (2023) tried to capture public opinion on climate change by utilizing two nationally representative climate change surveys. They found that LLMs (GPT-3) could effectively capture presidential voting behaviors to decide the climate change policies. Also, GPT-4 exhibits improved performance when conditioned on both demographics and covariates. Several studies used LLMs as a natural language processing approach to investigate public attitudes and sentiment analysis (Koonchanok *et al.*2023) for public attitudes of news topic by Twitter, (Ahmad*et al.* 2024) for COVID-19 vaccination sentiment analysis).

LLMs have the potential to be exploited for creating fake news, spreading propaganda, and manipulating public opinion (Chen & Shu 2023), while their capacity to process extensive volumes of text also enables the identification of instances involving fake news, propaganda, and the manipulation of public opinion. This analytical capability proves crucial in illuminating the narratives being disseminated and their potential impact on public perception and policy-making. Recognizing the influence of false narratives, particularly concerning issues like climate change, becomes imperative to prevent a lack of action or misguided policies. However, Patil *et al.* (2024b) provided the LLM approach for recognizing false news using the LamaIndex (https://www.llamaindex.ai). So, multi-modal LLMs using LLM with false-news detected LLM would be useful to summarize the public opinion.

A significant advantage of utilizing LLMs for such analyses lies in their immediacy. While a systematic review stands as a foundational method for summarizing knowledge on a particular subject, it demands considerable effort and essentially overlooks contemporaneous sources such as news articles and social media content. Conversely, the analysis of recent reports and information can alleviate the inherent limitations of the systematic review approach. However, the application of LLMs in ecology and biodiversity conservation presents its share of challenges. The accuracy and reliability of LLMs hinge on the quality of the data used for their training. Moreover, LLMs have the potential to perpetuate biases and inequalities present in the data, posing a risk to the effectiveness of conservation management efforts. Responsible use of LLMs is paramount, considering not only their accuracy but also the environmental impact of deploying these models for conservation activities. LLMs should complement human expertise in conservation rather than replacing it, acknowledging the unique insights, intuition, and contextual understanding that models may not fully grasp. The incorporation of LLMs in ecology and conservation necessitates navigating intricate legal landscapes, ensuring compliance with various international and national regulations.

Development of Ecology-specialized LPMs

Recently, there have been developments in specialized LPMs for specific purposes (Ling *et al.* 2023), such as specialized LLMs, BioMedLM, and agriGPT, which focus on biomedicine and agriculture. It is worth considering the potential benefits of developing specialized LLMs in ecology and biodiversity science, as they could potentially accelerate the advancement of the field. Specialized LPMs in ecology have the potential to enhance our understanding of complex ecological systems and contribute to decision-making processes related to biodiversity and ecosystem conservation. Developing such a model requires a multifaceted approach to create an AI system that can comprehend, generate, and analyze ecological data and language at a high level of sophistication. The development of LLMs specialized in Ecology is considered a significant advancement

in ecological research and biodiversity conservation.

These models, when combined with LVM offer unparalleled opportunities to revolutionize our understanding of complex ecological systems. There are some models for specializing LLMs such as knowledge-updated domain specialization (KUDS) with fine-tune step (Ling *et al.* 2023). KUDS needs more cost and less real-time use with tuning the LLMs, but for ecological knowledge based on long-term history would be suitable to use KUDS for the tuning. Through the use of LLMs with KUDS encompassing ecological literature and field observations, researchers can leverage their significant computational power to reveal complex relationships, forecast ecosystem dynamics, and provide support for conservation strategies based on evidence. The integration of LVMs and LLM-LPM into ecological studies enables the synthesis of diverse data sources, including satellite imagery, remote sensing data, and environmental sensor networks, thereby providing comprehensive insights into ecosystem health and resilience. In fact, (Jain *et al.* 2023) developed a LLM-LVM for text generation such as landscape category using satellite imagery. The combination of Ecology-specialized LPMs, including LLM, LVM, and multi-modal LPM, has the potential to advance ecological research and promote interdisciplinary collaboration. This could lead to more effective conservation initiatives aimed at protecting biodiversity and ecosystem services on a global scale.

Negative impacts of using LPMs on the environment

Rillig *et al.* (2023) noted that the energy consumption of LLMs during training and operation has significant adverse effects on the environment. The environmental impact of LLMs is closely linked to their immense computational requirements, which translate into extensive electricity usage. For instance, the training process of GPT-3 alone has been estimated to consume over 1287 MWh of energy (Rillig *et al.*2023). According to the studies conducted by Strubell *et al.*(2019) and Patterson *et al.* (2021), the estimated carbon footprint associated with this is 552.1 t of CO_2. It is worth noting that as LLMs continue to evolve and their demand for larger datasets increases, their energy requirements are expected to escalate even further (e.g., Luccioni *et al.* 2023). Therefore, it is important to evaluate and address the environmental impact of LPMs in a conscientious manner. It is crucial to align with responsible environmental practices and emphasize the necessity of ethical and eco-friendly advancements in the field of LPMs.

Conclusion

We examine here the potential of LPMs to address challenges in ecology and biodiversity conservation. The development and integration of specialized LPMs, particularly those tailored for ecology and biodiversity conservation, represent a significant advancement in ecological research and conservation practices. Moreover, multi-modal LPMs with LVM have the potential to offer extensive insights into ecology and biodiversity conservation by integrating various data sources and texts in different languages. LPMs are considered to substantially improve methods for ecology and biodiversity conservation. Nonetheless, it is imperative to collaborate among researchers, practitioners, and stakeholders to achieve a successful integration of LPM applications in ecological research and conservation practices.

Acknowledgments

We would like to thank Akira S. Mori for his comments on our manuscript. DeepL Write has been valuable in helping us to improve our English writing of this paper.

References

Ahmad, I.S., Aliyu, L.J., Khalid, A.A., Aliyu, S.M., Muhammad, S.H., Abdulmumin, I., *et al.* (2024). Analyzing COVID-19 Vaccination Sentiments in Nigerian Cyberspace: Insights from a Manually Annotated Twitter Dataset. *arXiv [cs.CL]* .

Albrecht, J., Kitanidis, E. & Fetterman, A.J. (2022). Despite "super-human" performance, current LLMs are unsuited for decisions about ethics and safety. *arXiv [cs.CL]* .

Amano, T., Berdejo-Espinola, V., Akasaka, M., de Andrade Junior, M.A.U., Blaise, N., Checco, J., *et al.* (2023). The role of non-English-language science in informing national biodiversity assessments. *Nature Sustainability* , 6, 845–854.

Amano, T., Berdejo-Espinola, V., Christie, A.P., Willott, K., Akasaka, M., Báldi, A., *et al.* (2021). Tapping into non-English-language science for the conservation of global biodiversity. *PLoS Biol.* , 19, e3001296.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P.,*et al.* (2020). Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.* , 33, 1877–1901.

Chen, C. & Shu, K. (2023). Combating Misinformation in the Age of LLMs: Opportunities and Challenges. *arXiv [cs.CY]* .

Cheng, W., Liu, J., Deng, J. & Ren, F. (2024). SSP: A Simple and Safe automatic Prompt engineering method towards realistic image synthesis on LVM. *arXiv [cs.CV]* .

Gast, R., Knösche, T.R. & Kennedy, A. (2023). PyRates-A code-generation tool for modeling dynamical systems in biology and beyond. *PLoS Comput. Biol.* , 19, e1011761.

Hannah, K., Haddaway, N.R., Fuller, R.A. & Amano, T. (2024). Language inclusion in ecological systematic reviews and maps: Barriers and perspectives. *Res Synth Methods* .

Jain, H., Verma, S. & Gupta, S. (2023). Investigating Large Vision Model Training Challenges on Satellite Datasets. In: *InGARSS 2023 - India Geoscience and Remote Sensing Symposium* .

Koonchanok, R., Pan, Y. & Jang, H. (2023). Tracking public attitudes toward ChatGPT on Twitter using sentiment analysis and topic modeling.*arXiv [cs.CL]* .

Lee, S., Peng, T.Q., Goldberg, M.H., Rosenthal, S.A., Kotcher, J.E., Maibach, E.W., *et al.* (2023). Can Large Language Models Capture Public Opinion about Global Warming? An Empirical Assessment of Algorithmic Fidelity and Bias. *arXiv [cs.AI]* .

Ling, C., Zhao, X., Lu, J., Deng, C., Zheng, C., Wang, J., *et al.*(2023). Domain Specialization as the Key to Make Large Language Models Disruptive: A Comprehensive Survey. *arXiv [cs.CL]* .

Luccioni, A.S., Viguier, S. & Ligozat, A.-L. (2023). Estimating the Carbon Footprint of BLOOM, a 176B Parameter Language Model. *J. Mach. Learn. Res.* , 24, 1–15.

Maurer, G.E., Hallmark, A.J., Brown, R.F., Sala, O.E. & Collins, S.L. (2020). Sensitivity of primary production to precipitation across the United States. *Ecol. Lett.* , 23, 527–536.

Parashar, S., Lin, Z., Li, Y. & Kong, S. (2023). Prompting Scientific Names for Zero-Shot Species Recognition. *arXiv [cs.CV]* .

Patil, D.D., Dhotre, D.R., Gawande, G.S., Mate, D.S., Shelke, M.V. & Bhoye, T.S. (2024a). Transformative Trends in Generative AI: Harnessing Large Language Models for Natural Language Understanding and Generation.*Int J Intell Syst Appl Eng* , 12, 309–319.

Patil, M., Yadav, H., Gawali, M., Suryawanshi, J., Patil, J., Yeole, A.,*et al.* (2024b). A Novel Approach to Fake News Detection Using Generative AI. *Int J Intell Syst Appl Eng* , 12, 343–354.

Patterson, D., Gonzalez, J., Le, Q., Liang, C., Munguia, L.-M., Rothchild, D., *et al.* (2021). Carbon Emissions and Large Neural Network Training. *arXiv [cs.LG]* .

Rane, N.L., Tawde, A., Choudhary, S.P. & Rane, J. (2023). Contribution and performance of ChatGPT and other Large Language Models (LLM) for scientific and research advancements: a double-edged sword.*International Research Journal of Modernization in Engineering Technology and Science* .

Rillig, M.C., Ågerstrand, M., Bi, M., Gould, K.A. & Sauerland, U. (2023). Risks and Benefits of Large Language Models for the Environment.*Environ. Sci. Technol.* , 57, 3464–3466.

Shah, D.S., Schwartz, H.A. & Hovy, D. (2020). Predictive Biases in Natural Language Processing Models: A Conceptual Framework and Overview. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* . Association for Computational Linguistics, Online, pp. 5248–5264.

Strubell, E., Ganesh, A. & McCallum, A. (2019). Energy and Policy Considerations for Deep Learning in NLP. *arXiv [cs.CL]* .

Wang, H., Liu, H., Cao, G., Ma, Z., Li, Y., Zhang, F., *et al.*(2020). Alpine grassland plants grow earlier and faster but biomass remains unchanged over 35 years of climate change. *Ecol. Lett.* , 23, 701–710.

Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.-S., *et al.* (2021). Ethical and social risks of harm from Language Models. *arXiv [cs.CL]* .

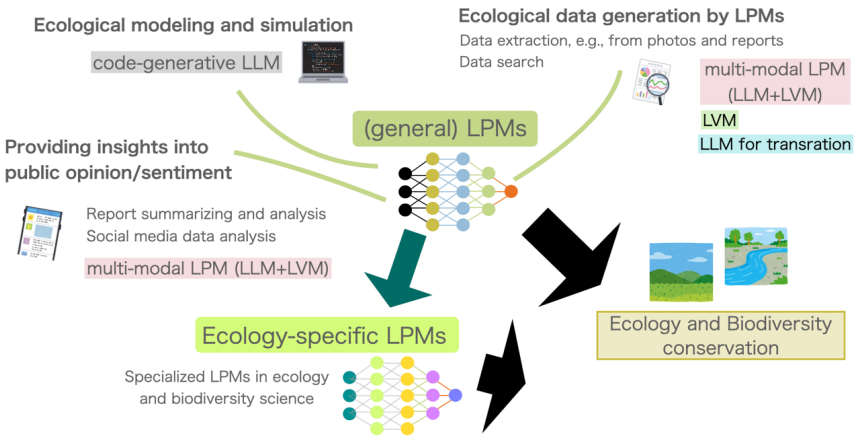

Figure 1 The illustration of the key applications of LPMs and Ecology-specialized LPMs for ecology and biodiversity conservation.